Cost Estimation Tool Set for NASA's Strategic Evolution of ESE Data Systems

Vanessa GRIFFIN[1], Kathleen FONTAINE[1],
Gregory HUNOLT[2], Arthur BOOTH[2], David TORREALBA[2]

[1]NASA / Goddard Space Flight Center, Greenbelt, MD 20771

[2]SGT, Inc., 7701 Greenbelt Rd, Greenbelt, MD 20770

Vanessa.L.Griffin.1@gsfc.nasa.gov, Kathleen.S.Fontaine.1@gsfc.nasa.gov, ghunolt@excel.net, abooth@sgt-inc.com, torrealba@sgt-inc.com

Abstract -- NASA's Earth Science Enterprise (ESE) is planning for the evolution of our existing data systems and data centers over the next 6-10 years. While recognizing the success of the EOSDIS and other existing data systems, we need a strategy for responding to the changing science requirements and for adapting to technology changes. A key component of the Strategic Evolution of ESE Data Systems (SEEDS) is to realize a far more heterogeneous and distributed system of data service providers than we have currently.

A key facet in the planning for SEEDS is to create data service cost estimation tools for use by principal investigators proposing to future research announcement and suitable for Program Office estimation of the overall costs for various architecture and implementation options. The cost estimation tools are based on "costing by analogy" using a database of data management costs for previous science research missions. We are also using commercial cost estimation tools to estimate the development costs for new data systems. The paper will discuss the status of the tool development, the approach taken for enterprise cost estimation, and lessons learned from development of the SEEDS cost estimation tools and linking various cost modeling tools into an integrated tool set.

1. Introduction

NASA's Earth Science Enterprise (ESE) presently operates a distributed data and information system to collect, process, catalog, archive, distribute Earth science data and products to users and provide support to users. While the immediate users are scientists and applications specialists participating in the ESE science and applications program, ESE data and products are available to and widely used by the general scientific, educational and applications community. Today's distributed ESE data and information system includes nearly seventy data systems, including eight Distributed Active Archive Centers (DAACs), thirty Earth Science Information Partners (ESIPs), and numerous others.

Over the next seven to ten years NASA / ESE will be launching nine new flight projects to collect data needed for the study of the Earth system. The research and applications program will add many new activities that either through the increasingly interdisciplinary nature of the research, or the need to couple applications and research, place increasing demand for interoperability between elements of the distributed data and information system. New flight projects will be creating larger volumes of new data and products, and the research into climate change and the Earth system will require analysis of longer and longer time series of data and products from multiple sources. At the same time, tight budgets will compel ESE to maximize the efficiency of its data and information services by building on the most successful aspects of current capabilities and making the best possible use of new but proven technology.

The figure below illustrates conceptually the expected nature of the distributed ESE data services architecture, illustrating the interplay between three types of ESE data services providers, Mission Data Systems that support flight projects, Science Data Centers that support specific research efforts, and Backbone Data Centers that a robust underpinning to the ESE data services architecture, data services to the broad user community, and preservation of ESE data until it is transferred to the U.S. agencies responsible for long term archiving.

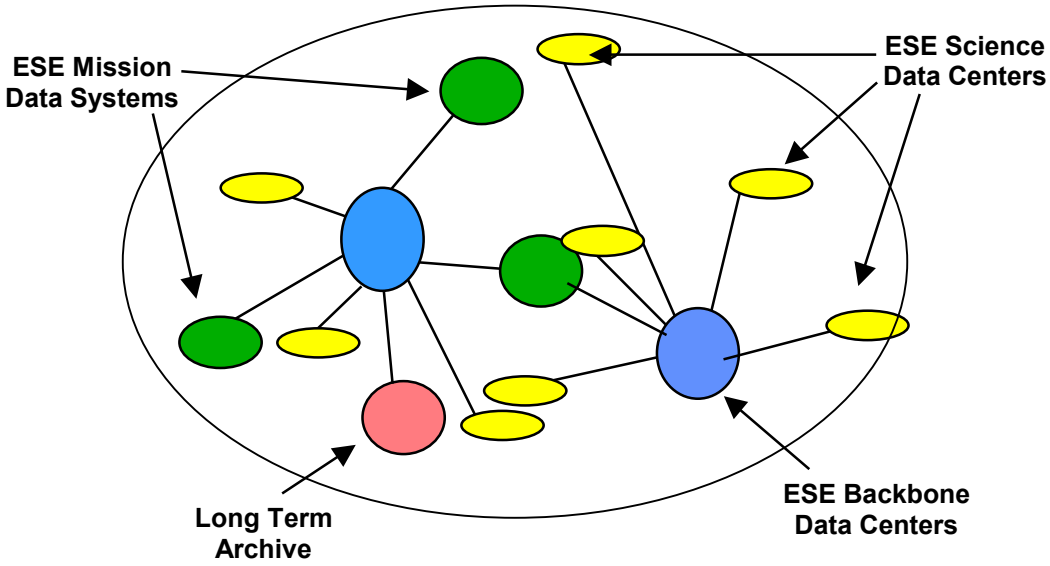

Figure 1 - A schematic view of the future ESE data services architecture.

In August of 1998, the NASA ESE Associate Administrator directed the development of a plan for evolution of the current ESE distributed data and information system to best meet ESE's data management and user services needs over the next seven to ten years. Originally called a plan for a 'New Data and Information Services System, or NewDISS', this is now known as a plan for the Strategic Evolution of the ESE Data System, or SEEDS. After a concept study was completed in 2000, a SEEDS Formulation Team was established in 2001 at NASA's Goddard Space Flight Center to develop an implementation approach for the evolution of the ESE data and information system.

2. Need for Cost Estimation

One of the key questions facing the ESE and the SEEDS Formulation Team is how to estimate the costs for implementing and operating elements of the ESE data and information system over the next seven to ten years. This requires a companion analysis into the requirements that the ESE data and information system must meet, requirements that flow from ESE's science and applications objectives and are the basis for the size and scope of the capabilities needed by ESE.

An ability to estimate costs is needed from two different user perspectives—the individual investigator or PI (principal investigator), and the ESE program office. Individual investigators and projects within the ESE need an ability to estimate costs for specific data service activities that they will propose as part of their response to a relevant ESE solicitation.

The ESE program office needs to have the best possible understanding of the overall costs of the ESE data services architecture (i.e., the set of interoperating elements and connections between them). There are many possible ESE data service architectures (see figure above), reflecting many different possible allocations of ESE data management roles and missions to existing or new data service providers. The ESE program office must have the ability to estimate the cost of different data service architectures if it is to be able to make intelligent cost-benefit trade-offs between possible approaches.

The SEEDS Formulation Team began, in October, 2001, a two year effort to develop a baseline set of requirements and levels of service, and a cost estimation capability for use by the ESE program office and by individual investigations and projects. That effort, the Levels of Service / Cost Estimation (LOS/CE) study, is the subject of this paper.

3. Cost Estimation Approach

The cost estimation tool being developed by the LOS/CE study will be based on a "cost estimation by analogy" model. The model will base its life cycle cost estimates for future activities on a ‘comparables database’ of information describing existing or very recent data management activities seen as functionally analogous, or ‘comparable’, to future activities. Underpinning both the cost estimation by analogy model and the comparables database is a generic data service provider reference model developed to establish a framework for coupling requirements / levels of service with costs through a set of parameters that include cost estimation outputs, comparables database parameters, and parameters that a user would provide to describe an activity to be estimated. All of these parameters are organized by the functional areas / areas of cost that comprise the generic data service provider.

3.1 Data Service Provider Reference Model as Framework for Cost Estimation

The cost estimation by analogy model and comparables database are structured around a generic data services provider reference model. The reference model has three related elements:

1) A set of ‘functional areas’ that collectively comprise the full range of functions that a data service provider might perform and the areas of cost that must be considered by the cost estimation by analogy model. These include ingest, product generation, archive, search and order (or automated equivalent), access and distribution, user support, etc.

2) A set of parameters for each functional area that constitutes a quantitative description of the workload, staff effort, and any other factors that contribute to cost for that area. These parameters are common to all data service providers as they apply.

3) A set of requirements and levels of service for each functional area.

These three aspects of the model are closely coupled to ensure the internal consistency of the model. The set of functional areas is the underpinning; both the model parameters and requirements / levels of service are organized according to the functional areas. The requirements / levels of service and the model parameters are coupled in that the definitions of the requirements / levels of service embody model parameters. This integration of the three elements of the model is intended to ensure that estimated costs are driven by and traceable to requirements to the fullest extent possible.

The general data service provider reference model includes all functions / areas of cost that a generic data service provider might perform. While an actual working data service provider could conceivably perform all of the functions included in the model, most if not all actual data service providers perform a subset of them.

3.2 Cost Estimation by Analogy

The cost estimation by analogy technique is based on the assumption that reasonably reliable estimates for the cost of a future data activity (either by a new organization formed for that purpose or as an increment to the data activities of an existing organization, or some combination of these) can be based on an analysis of the past history and experience with other similar data activities. This assumes that a sufficient sample of reasonably applicable cases exists on which to base an estimate, and that present-day, recent past cases are applicable in implementation and operation approach as well as function and workload, so that the effort required for the cases can be taken as suggestive of the effort that will be required for a new case to be estimated.

The first assumption is important when statistical techniques such as regression are considered; if there is too small a sample the results will be unreliable or entirely useless, as will be indicated by the probable errors of estimate that will accompany the estimates, and by the results of tests on independent cases.

The second assumption reflects the concern that a project that might be nearly identical in terms of the nature of the data activity (function and workload) to be estimated but might have been done (implemented and/or operated) by an approach so different as to compromise partly or completely its value as a data point for producing an estimate for a new activity. Attention must be paid to trends that could follow changes in approach that might provide a basis for an extrapolation into the future.

The output of the cost estimation by analogy model (or the tool which embodies it) is a set of parameters, a year by year spread of selected cost factors and supporting information (e.g. staffing). The model employs a set of 'cost estimating relationships' (CERs) to compute its output from the input provided by the user. The model employs three kinds of cost estimating relationships (CERs) - 'plug value', 'arithmetic', and 'statistical'. The first kind of CER is the 'plug value'. Plug values are constants used when there is no better way of computing the output parameter. The second kind of CER is 'arithmetic'. In this case there is a simple arithmetic relationship between the output and its input(s). The third kind of CER is 'statistical'. The output parameter is computed by a relationship that involves the data (i.e. one or more parameters) from the comparables database. The relationship may be based on linear regression, a non-linear relationship, or other statistical techniques. An error of estimate will accompany the result. In addition, commercial cost estimation tools (e.g. based on the Constructive Cost Model - COCOMO) will also be used to estimate the development costs for new data systems. The net result is a cost estimation by analogy supplemented by parametric techniques.

The information in the comparables database (though assembled on a site by site basis) will be used on a parameter by parameter basis within the reference model's functional areas. The 'best fits' for a projected new data activity's ingest area might includes cases that were not good fits for other areas, etc.

The cost estimation by analogy model will not directly estimate future costs on the basis of past costs. It is indeed almost a misnomer to call the model a 'cost model' because the real basis for comparison with cases is staff effort and system capabilities. Year by year costs are only added as a final step. A year by year effort estimate is first produced, and then priced out by application of labor rates and inflation. Similarly projections of required system capabilities are made, and then priced out through use of system capability vs projected cost curves. Other non-staff elements of cost are handled in like manner. Finally all factors are summed to produce the final output, the year by year life cycle cost estimate.

4. Development Approach

The general approach taken by the LOS/CE study team to the development of the cost estimating capability is to work "top down", to begin with a working model that demonstrates how the model will run, the user interface, and the output to be produced, based on a simple set of cost estimating relationships, even dummy placeholders if necessary.

The cost estimation by analogy model will evolve over the life of the project, based on feedback from users evaluating prototypes and eventually actual experience with use of the model, and based on the development of the comparables database. The development of the comparables database will drive the evolution of the model because the CERs used by the model are dependent on, or constrained by, the state of the comparables database. The CERs have to be developed through analysis of the available data. The state of the available data will develop slowly as the information collection process goes on - i.e. as the comparables database is gradually built. In the case of some parameters, a sufficient number of comparable cases will be accumulated to enable statistical relationships to be used. For other parameters this will not be the case, and either reasonable arithmetic approximations will be used or the parameters will have to be dropped. The model has to be flexible to accommodate inevitable changes to the CERs as more is learned about the available data and as various possible combinations of parameters are tested to see which combinations yield the strongest relationships. At first only simple relationships will be employed, but as development proceeds the use of non-linear relationships will be explored, and perhaps tools / techniques that evaluate the relative 'distance' of the input case to the members of the set of comparables to produce a better estimate. As the model is developed and as the CERs are refined, the model will be tested against independent data for actual data service providers not included in the comparables database (for whom the actual outputs are known).

5. Status and Plans

An initial definition of what information is needed as output from the cost estimation model has been developed - this describes the specific content of a life cycle cost estimate for a data activity, year-by-year costs in various categories with selected supporting information (e.g. staffing levels).

An initial version of a data service provider reference model has passed through a first round of review. The model's functional areas and requirements and levels of service have been updated recently and further

feedback is being sought. A working definition of the parameters that make up the model has been completed. A preliminary description of the relationships between the parameters in the form of sets of input-process-output relationships has been defined, at this point with placeholders for the process steps. These placeholders will be replaced by progressively more refined CERs as the project proceeds.

A major effort has begun to collect the information from existing data activities that is needed to build the comparables database. The effort to collect information and build the comparables database will proceed for many months. The near term intent is to get a sufficient sample to support model development and a demonstration prototype capability by October, 2002.

The demonstration prototype will be a 'proof of concept'. It will show how the cost estimation tool will work, how a user will use it, how the ESE level and investigator / project level scenarios will be realized. The demonstration prototype will use a very limited comparables database. It will employ an initial set of simple CERs, some linear equations for comparables-based CERs supplemented by simple parameterization as needed (i.e., 'plug-value' or 'arithmetic' relationships as described above). It will produce a life cycle cost estimate, regardless of what simplifications are necessary at this point. Its ability to produce useful results will be constrained by the small size of the comparables database. The ability to produce useful results depends on the database of comparables being as large as possible, allowing the best CERs, and in the demonstration prototype timeframe the comparables database will not contain enough cases. Results will not have been tested against independent cases - that will come later when more data is collected and some cases can be held aside for such testing.

The demonstration prototype will be iteratively refined based on user feedback, the development of better CERs as the comparables database is built, etc., with a sequence of more refined prototypes culminating in an operational capability by September, 2003.

Although a discussion of 'lessons learned' from the cost estimation effort is premature at this point, a key area of caution can be highlighted. The reliability of the life cycle cost estimates produced by the tool being developed will depend on the size and quality of the comparables database. The size of the sample (of data service providers represented in the comparables database) is likely to be marginal for statistical significance, especially if information about non-NASA U.S. data activities and international data activities can not be obtained. While the estimates could reveal trends and provide rough indication of costs, the output of the model might not be suitable as the sole or primary basis for proposal submission numbers.

An advantage of the model will be that it will permit the costs of effort essential for data preservation to be planned in to all elements of the ESE data services architecture as they apply. For example, the model will help activities (such as the current DAACs) that hold data for significant periods of time plan for archive quality media, monitoring and refresh of archive media, etc. The model will also help activities (such as flight project data systems or science data centers) that primarily produce data and products plan for development of documentation of the quality needed for long term use of their products.

The LOS/CE study will not be successful without feedback and guidance from the community of data users and data service providers, including comments given at workshops, and review of project documents, and will include evaluation of prototypes. Once an operating version of the cost estimation capability is generated, continued use will enable iterations for improved prediction capability.

6. Data Center Benchmark Study

The general data service provider reference model described above is an extension of a reference model developed in the course of a comparative analysis of data center operations as of the year 2000 that was performed by the ESDIS Project in 2000 and 2001. (The study report, "ESDIS Data Center Best Practices and Benchmark Report", by G. Hunolt and A. Booth, SGT, Inc., on contract to NASA, is available on request on CD-ROM). The current LOS/CE study in essence extends the scope of the reference model to cover the full life-cycle of the data service providers (including initial implementation) and adds the capability to estimate future effort and therefore costs. The approaches developed in the course of the benchmark study to normalization of information across different sites to enable meaningful comparisons to be made will be used by the current LOS/CE study in its cost estimation by analogy model. For these reasons a look at the benchmark study and its results is relevant.

The goal of the previous study was to assess the reasonableness of the staffing levels of the EOSDIS DAACs for the work they do, and, if possible to identify areas where there may be potential to improve cost effectiveness". This required development of an understanding of the relationships between data center operations workload and the staff effort required to accomplish it, and an understanding of what 'best practices' of other data centers could be useful lessons for NASA's data centers, especially to increase the cost effectiveness of operations. Information was collected from three NASA DAACs and eleven other data centers, including three from Europe as well as U.S. NASA and NOAA data centers (their cooperation was greatly appreciated).

In order to allow a comparative analysis, the information received from the different data centers had to be reconciled to a common framework, and the original data service provider reference model was developed for that purpose. It included a set of operations functional areas and a set of parameters for each, and information received from the cooperating data centers was mapped to the reference model, producing a basically consistent set of parameters across all of the sites. The mapping could not be perfect given the differences between the data centers and their own functional view and their own methods of measuring their own effort and workload.

It was also necessary to arrive at a few high level measures of workload and an approach to normalize for the differences in scale across the sites. The approach taken to normalization was to obtain annual workload measures (either by using annual measures directly or by extrapolating annual measures from measures taken over shorter intervals), simply compute workload measure per unit of effort, and then base comparisons on the resulting rates of work done per effort expended. These rates were interpreted as rough measures of productivity.

Three workload measures were included, representing approaches from two directions and a synthesis: 'volume of data managed', 'product traffic managed', and the synthesis, 'work'. The first measure, 'volume of data managed', is computed as the sum of the annual volume of data ingested, volume produced, 10% of the archive or working storage size, and the volume distributed. Work scales with the volumes of data involved with the different functions; more volume, more work. Including 10% of the archive volume (given that addition to and retrieval from the archive for distribution is already counted) was a rough reflection of archive maintenance or periodic refresh.

By itself the 'volume of data managed' metric misses the complexity of the work that follows from the structure of the data handled, whatever its volume. Handling a large number of individually small data products can be more work than handling a small number of very large data products. To measure work from a product point of view, the second parameter, 'product traffic managed', was calculated as the sum of the annualized number of products ingested, generated, and distributed.

Finally, both volume and product count contribute to the total work done. A third parameter was defined that attempted to reflect the contributions of both to work. This parameter, simply named 'work', is computed as the sum of 'volume of data managed' in gigabytes and 'product traffic managed' divided by 1000. The scaling of volume and factor of 1000 tend to balance the contribution of volume of data managed and the product traffic managed to the total 'work'. This is an arbitrary normalization that yields a reasonable spread of values. It is weighted toward volume, which seems reasonable since volume by itself slows product rates, potentially increasing operator time per event.

One final preliminary step was to group the sites by class according to their function. The three DAACs and eleven comparison data centers fell into three similar groupings: A - the three DAACs and four comparison data centers performed in all functional areas, B - two sites did all but product generation, and C - five sites were large scale product generation sites only. In the case of the DAACs it was necessary to make a further distinction between their internally developed systems and the externally developed and provided EOSDIS Core System (ECS). Each DAACs was treated as two sites, an "ECS" site and a "Non-ECS" site.

Table 1 below which presents volume of data managed vs. staff effort.

Table 1–Volume of Data Managed (in Terabytes) vs. Staff Effort

Volume of Data Managed vs. Staff Effort Metrics	All Classes Survey Site Average	Class A & B Survey Site Average	Class C Survey Site Average	DAAC Site Average	DAAC "ECS" Site Average	DAAC "Non-ECS" Site Average
Volume of Data Managed, TB	78.38	32.44	133.50	250.68	209.07	41.77
Volume of Data Managed, TB per FTE	1.9	0.8	4.3	2.4	3.1	1.1
Volume of Data Managed, TB per Ops FTE	4.3	2.6	10.4	7.7	13.3	2.5

In Table 1 the measures shown are volume of data managed, volume of data managed per FTE, and volume of data managed per operations FTE. Two items stand out. The Class A/B survey sites are comparable to DAAC "Non-ECS" 'sites' in terms of both the overall VDM and VDM per total and ops effort. The Class C sites are likewise comparable to DAAC "ECS" 'sites' in both overall VDM and VDM per total ops effort. The Class C sites and DAAC "ECS" 'sites' are alike in including large scale product generation functions, and the distribution functions are partly comparable, the DAAC "ECS" 'site' number including a large operational distribution to a SIPS. Still, the DAAC "ECS" site stands out with the best productivity value for VDM per ops FTE for a more complex function set than the Class C sites.

Table 2 presents the synthesis measure, 'work', 'work' per FTE and 'work' per operations FTE.

Table 2–' Work' vs. Staff Effort

Work vs. Staff Effort	All Classes Survey Site Average	Class A & B Survey Site Average	Class C Survey Site Average	DAAC Site Average	"ECS" Site Average	"Non-ECS" Site Average
Work = (VDM in GB + Products/1000)	90,291	34,297	155,473	256,225	211,474	44,751
Work per FTE	2,173	993	3,113	2,407	3,090	1,177
Work per Ops FTE	4,940	2,176	7,299	7,873	13,416	2,667

'Work' seems to reasonably reflect the size and scope of activity of the groups of sites, i.e. it seems reasonable that the DAACs on average do a bit more 'work' than the Class C sites that are larger scale but more limited in scope, and that the DAAC "Non-ECS" 'sites' are comparable but do a bit more work than the Class A/B survey sites which are similar but smaller.

What stands out are the comparisons of 'work' per total effort FTE and perhaps especially 'work' per ops FTE. The DAAC "ECS" 'site' stands out distinctly in 'work' per ops FTE, appearing to be much more productive in that sense than any other site group. The DAAC "ECS" 'site' also is equal in overall productivity to the Class C site average.

Three key results were produced by the study:

1. On the whole, the DAACs are at least on a par with the survey sites for overall workload productivity, and in most cases appear to be a bit better. A productivity advantage would translate to better cost effectiveness if unit labor rates and skill levels are roughly comparable. The DAAC staff effort, and by implication costs, are not out of line with the external data centers included in the study.

2. The ECS seemed to deliver significantly better productivity in 2000 than other large scale production systems, at a higher cost of maintenance.

3. DAAC locally developed "Non-ECS" systems collectively are on a par with similar locally developed and implemented systems at the survey sites.

The same approaches to workload measures and normalization will be used by the LOS/CE study in developing the comparables database and the cost estimation by analogy model. Information from the sites included in the benchmark and best practices study will be included in the comparables database if the one year snapshot can be extended to include the lifetime of the sites.

7. Summary

The NASA Earth Science Enterprise must enhance and extend its data management and services capabilities to support its aggressive science and applications program. In a time of tight budget constraints, success depends on making the best use of resources, on an individual project basis as well as across the ESE as a whole. This requires a capability to estimate life cycle costs for individual data activities as well as for an ESE architecture of data activities. This cost estimation capability will be based on current and recent past experience with ESE and other similar data activities. A 'comparables' database describing as many such activities as feasible is being compiled and will be updated and maintained. The cost estimation capability will use cost estimation relationships derived from the comparables database to estimate the effort required for a new data activity based on its mission and expected workload, arriving at cost based on the estimated effort and expected rates (e.g. labor rates). A data services provider reference model that describes the functional areas and areas of cost for a generic data services provider will provide the framework for the comparables database and the cost estimation capability. The model will be an extension of a reference model developed for an earlier study, and approaches for workload measures and normalization for scale developed for that study will be carried over into the current study.

The needs of the ESE science and applications program will evolve over time. Consequently the ESE roles and missions for data service providers that support the program will evolve. All the while, the information technology that touches all aspects of every data service provider and the user community will evolve. The data service provider reference model and the cost estimation tool will have to evolve accordingly to remain relevant and useful. They will be improved in successive iterations as the comparables database grows and includes more new and updated activities, and on the basis of lessons learned derived from use of earlier versions of the model.

The SEEDS Website, [http://eos.nasa.gov/seeds/], contains information about the SEEDS effort in general, all of the Formulation studies, and more detailed information concerning the LOS/CE study discussed in this paper, including a set of six LOS/CE working papers that discuss the data services provider reference model and the cost estimation by analogy approach.